ED 429 998 TM 029 708

AUTHOR	Dawadi, Bhaskar R.
TITLE	Robustness of the Polytomous IRT Model to Violations of the Unidimensionality Assumption.
PUB DATE	1999-04-00
NOTE	50p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; Estimation (Mathematics); Factor Analysis; *Item Response Theory; Models; *Robustness (Statistics); Simulation
IDENTIFIERS	*Polytomous Variables; Unidimensionality (Tests); *Violation of Assumptions

ABSTRACT

The robustness of the polytomous Item Response Theory (IRT) model to violations of the unidimensionality assumption was studied. A secondary purpose was to provide guidelines to practitioners to help in deciding whether to use an IRT model to analyze their data. In a simulation study, the unidimensionality assumption was deliberately violated by using two-dimensional data. The "impact," or change in the error due to violating the assumption, was calculated to assess the effects of the violation on ability estimation. The effects of problematic variables on absolute impacts and their interactions were analyzed, and a factor analysis using the principal component method was conducted to provide guidelines to practitioners on the computer-generated generated data. The precision of the estimated ability was determined in four ways. When the ability estimate was assumed to measure the average ability and the major ability of two unequally important abilities, the procedure was generally robust to the violation. However, when the ability estimate was assumed to measure one of the two equally important abilities and the minor ability of two unequally important abilities, the estimation procedure was not robust. Results from the analysis of variance were consistent with the results from analyzing the relative impacts. (Contains 5 tables and 47 references.) (Author/SLD)

Robustness of the Polytomous IRT Model to Violations of the Unidimensionality Assumption

Bhaskar R. Dawadi
Georgia Examining Boards

A paper presented at the annual meeting of AERA
Montreal, Canada, April 1999

Abstract

The primary purpose of this study was to investigate the robustness of the polytomous Item Response Theory (IRT) model to violations of the unidimensionality assumption. The secondary purpose was to provide guidelines to practitioners to help in deciding whether to use an IRT model to analyze their data.

The unidimensionality assumption was deliberately violated by using two-dimensional data. The "Impact," or change in the error due to violating the assumption, was calculated to assess the effects of the violation on ability estimation. Problematic variables were identified using Relative Impacts, and the effects of those variables on Absolute Impacts and their interactions were analyzed using a 2^6 factorial ANOVA. A factor analysis using the principle component (PC) method was conducted to provide guidelines to practitioners on the computer generated two-dimensional data.

The precision of estimated ability was determined in four ways by comparing the estimated ability: (1) with the average of two true abilities, (2) with one of the two equally important abilities, (3) with the major ability of two unequally important abilities, and (4) with the minor ability of two unequally important abilities.

When the ability estimate was assumed to measure the average ability (1) and the major ability of two unequally important abilities (3), the procedure was generally robust to the violation. However, when the ability estimate was assumed to measure one of the two equally important abilities (2) and the minor ability of two unequally important abilities (4), the estimation procedure was not robust. Results from ANOVA analysis were consistent with the results obtained from analyzing the Relative Impacts.

Introduction:

The demand for performance assessment is increasing in the field of educational measurement (Crocker, 1995). California has replaced the paper and pencil test with a new statewide examination using performance assessment ("California's New Academic Assessment System," 1996). Performance assessments allow scorers and assessors to evaluate skills achieved by students, skills that cannot be measured by traditional modes of evaluation such as a paper and pencil test (Oosterhof, 1990). A rating scale is commonly used to record observations in performance assessment. A polytomous item response theory (IRT) model is one of the appropriate tools for analyzing observations recorded by the rating scale (Andrich, 1978a and 1978b). IRT models have been used by test developers and measurement specialists in various applications, such as in customized testing, criterion-referenced testing, and national assessment (Hambleton, 1989).

Polytomous IRT models are based on a set of strong assumptions. The assumptions are necessary for the integrity of the models and these assumptions help to bring the mathematical complexity of the model within reasonable bounds. Unidimensionality is one of the assumptions of polytomous IRT models. The assumption of unidimensionality states that only one ability or trait is necessary to "explain" or "account" for an examinee's test performance (Hambleton & Swaminathan, 1985). However, in practice, data obtained from educational achievement tests do not always satisfy the unidimensionality assumption of IRT models (Traub, 1983). Constructs, even like vocabulary ability, are multidimensional if analyzed in enough detail (Reckase, Ackerman, and Carlson, 1988). Many studies have shown that when the unidimensionality assumption of the dichotomous IRT models was violated, the results obtained from such analyses were not valid (e.g., Folk & Green, 1989; Oshima and Miller, 1990; Dorans & Kingston, 1985; Downing and Haladyna, 1996).

Studies have been conducted to detect the effect of various degrees of violation of the unidimensionality assumption for the dichotomous IRT model (e.g., Reckase, 1979; Drasgow and Parson, 1987; Harrison, 1988; Zeng, 1989; Dirir and Sinclair, 1996). However, studies indicating the robustness of a unidimensional polytomous model when the test is multidimensional are rare thus far. To date and to the best of my knowledge, only DeAyala (1995), using computer simulated multidimensional data, has explored the systematic effect of multidimensionality on the estimation in the Master's partial credit model (Master 1982). This model assumes that all items are equally effective at discriminating among examinees. Parameter estimates were obtained in the study using the computer program MSTEPS (Wright, Congdon, and Schultz, 1989). The study was conducted using the compensatory and non-compensatory multidimensional data.

In a compensatory model with multidimensional data, the author found that the estimated ability ($\hat{\theta}$) was a better estimate of the mean ($\bar{\theta}$) of two abilities θ_1 and θ_2 than either individual ability θ_1 or θ_2. As the multidimensionality of the data decreased, the "differences in RMSE and bias with respect to θ_1, θ_2, and mean $\bar{\theta}$ diminished (p. 413)." In a noncompensatory model, when the data were multidimensional, the accuracy of estimated ability ($\hat{\theta}$) was consistently greater for mean ability ($\bar{\theta}$) than for either θ_1 or θ_2.. However, when the test was divided into two individual dimensions, the Root Mean Square Error (RMSE) corresponding to one of the examinee's abilities was lower than the RMSE with respect to the mean ability ($\bar{\theta}$). As the correlations between two abilities (θ_1 and θ_2) increased, the RMSE decreased; however, one ability (θ_1) was better estimated than the other ability (θ_2) throughout the θ continuum.

Purpose:

The primary purpose of this study was to investigate the robustness of the unidimensional polytomous IRT model when the data analyzed by the model were multidimensional. Data were generated to fit the generalized logistic partial credit model (GPCM) (Muraki, 1992a). The GPCM is derived by adding the *a* parameter (slope) in the equation for Masters' partial credit model. Thus, in this GPC model each item had different discriminating capability.

The secondary purpose of this study was to provide a general guideline to practitioners to help in deciding whether using an IRT model to analyze their data would be appropriate or not. In a simulation study like this, all results are obtained in terms of true parameters. However, in practice, true parameters of the data are not known to practitioners. Thus, a factor analysis using the principle component method (PCA) was used to generate such guidelines.

Method:

The study was conducted using computer generated data. The sample size was fixed to 1000 for all simulations in this study. In order to control the complexity of this study, only two dimensions θ_1 and θ_2 were used while generating item response data, i.e., the study included cases with violations of the unidimensionality assumption due to two underlying abilities (θ_1 and θ_2). Variables (or parameters) and their two levels systematically varied in the study are listed below in Table 1.

InsertTable 1 about here

Out of these six variables which were systematically varied during the study, all variables except Dimensional Strength were self explanatory and commonly used in this type of simulation study. In this study the variable Dimensional Strength was used to show the relative importance or strength of dimensions on a test, and it was done by increasing or decreasing the number of items representing each dimension. In a test with 8 items, if six items represent dimension one (θ_1) and two items represent dimension two (θ_2), common sense tells us that the Dimensional Strength (or the relative importance) of ability θ_1 is greater than that for the ability θ_2. Two levels of dimensional strength expressed by the proportion of items that represented each dimension were 50/50 and 25/75 percentages. On the 50/50 percentage strength level, each dimension was equally represented, whereas on the 25/75 percentage strength level, 25 percent of the total number of items in a test represented dimension-one and the remaining 75 percent of items represented dimension two.

Root Mean Square Error (RMSE) was calculated using true and estimated abilities for both unidimensional and two-dimensional data, separately. RMSE was calculated using the following formula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_{est} - \Theta_{true})^2}{n}}$$

where, x_{est} = individually estimated ability (θ) by PARSCALE for both uni and two-dimensional data

θ_{true} = true ability which is θ_1, θ_2, and the average ($\bar{\theta}$) of θ_1 and θ_2 for multidimensional data and simply θ for unidimensional data n = 1000, number of individuals taking the test

Data Generation:

Both unidimensional and two-dimensional data were generated using the RESGEN program (Muraki, 1996). Abilities in the unidimensional data were generated independently of the abilities from the two-dimensional data, i.e., unidimensional θ was not the same as θ_1 or θ_2 of two-dimensional data. However, abilities for both two-dimensional and unidimensional data were generated under the same condition.

The effect of a violation of the assumption of unidimensionality was assessed using true and estimated abilities. Estimation of abilities came from PARSCALE (Muraki and Bock, 1993). RMSE on both multidimensional and unidimensional abilities was calculated. For the unidimensional data, there was only one RMSE calculated using a unidimensional true ability (θ) and estimated ability ($\hat{\theta}$). For the two-dimensional data, three different RMSEs were calculated for each set of conditions using the single estimated ability ($\hat{\theta}$) and true ability θ_1; $\hat{\theta}$ and true ability θ_2; and $\hat{\theta}$ and the mean ability ($\bar{\theta}$), which was the average of true ability θ_1 and true ability θ_2, individually. PARSCALE did not provide the estimate of the average ability or individual ability. Only one ability estimate was obtained from PARSCALE for the two-dimensional data. For the simplicity of discussion in this study, the ability estimate obtained from PARSCALE was assumed to measure the average of the two true abilities and two individual abilities.

Analysis

Identifying Problematic Conditions:

The Impact of violation represented by Y was assessed by calculating the difference between the RMSE of the two-dimensional data ($RMSE_{2D}$) and the RMSE of the unidimensional data ($RMSE_{1D}$) , i.e., $Y = RMSE_{2D} - RMSE_{1D}$. If the Impact (or Y) were minimal, the model could be said to be robust. If the Impact were large, the model could be said to be not robust. During the study, Y was represented by IMPCT_AV, IMPCT_1 and IMPCT_2 for the average ability ($\bar{\theta}$), ability one (θ_1) and ability two (θ_2), respectively. For analysis purposes, the Impact was defined in two ways: Absolute Impact and Relative Impact.

The Relative Impact was the percentage increase of root mean square error of two-dimensional ($RMSE_{2D}$) data with respect to the root mean square error of unidimensional data ($RMSE_{1D}$). It was calculated, for example for IMPCT_AV, by dividing the mean of IMPCT_AV by $RMSE_{1D}$ and multiplying that dividend by 100. The RMSE for both unidimensional and two-dimensional data was calculated with 1000 subjects. The sample size of 1000 was enough to bring the variation due to chance between runs to a negligible minimum.

To give readers a sense of the practical importance of these Impacts, a threshold value of Relative Impacts was selected. It was assumed that a Relative Impact higher than 20 percent was practically important in this study. That is, an error of 20 percent or larger was regarded as a considerable amount of error for the purpose of identifying conditions in this study that were problematic to the violation. Any treatment condition that yielded a Relative Impact lower than 20 percent would be considered robust to violation of the unidimensionality assumption. For the discussion purpose, the treatment conditions based on the threshold value of Relative Impacts were further classified into three criteria: treatment conditions that were generally robust, generally not robust, and most problematic to violation of the assumption. Using Impact as the outcome allowed the researcher to answer the following two basic questions: 1) Under what conditions would the violation become problematic? And 2) What were the effects of each of the study variables on Absolute Impact ?

Developing Practice Guidelines Using the PCA Results:

To assist practitioners in using the IRT models for analyzing their data, results from the IRT analysis were compared with the results from the factor analysis. The guidelines were based on what an analyst would reasonably know about the data based on a simple factor analysis. From the results of the factor analysis, practitioners would know the number of existing factors in their data (i.e., dimension of ability), the correlations between those factors (if there were more than one factor), and the dimensional strength (or relative importance) of their data. The dimensional strength would be known by looking at the loadings of factors on each item. Analysts would also know from the outset the number of response categories of each test item and the length of test used. Thus, only those variables that were easily known to practitioners (the number of factors (i.e., dimension of ability), the correlations between factors (COR), the dimensional strength (DS), the number of response categories (CAT) and test length (TL)) were used to provide guidelines to practitioners in deciding for themselves whether to use an IRT model for data analysis, or not. Item Category Threshold Range (THR) and Item Slopes (SLP) were two variables out of six that were systematically varied in this study. However, those variables were not available to practitioners from the results of factor analysis because THR and SLP are specific to IRT analysis. Thus, THR and SLP were excluded when guidelines were developed for practitioners.

While developing the guidelines, first, the number of factors were determined for all treatment conditions. Second, Relative Impacts (already calculated) of the treatment conditions were compared with the threshold of 20 percent to determine the problematic conditions. Third, based on the correlations obtained from the factor analysis, guidelines to practitioners were developed. When only one factor was extracted, there was no correlation; thus, conditions other than correlation such as, the number of reponse categories and the dimensional strength, were applied to develop the guidelines. The reliability of these guidelines was judged by calculating the success rate (how many times the results obtained by using the guidelines were correct).

Effects of Study Variables on Impacts:

A 2^6 factorial design was used to test the significant effects of six variables and their interactions on Impacts. The Absolute Impact was used as the dependent variable in the analysis of variance (ANOVA). With six variables in the design and each factor having two levels, the analysis had 64 treatment conditions, and each run was replicated twice. This replication would provide variability within each cell. In the ANOVA analysis, the statistical significance of all main and interaction effects were tested, and the highest order interactions were identified. The practical importance of these higher order interactions was determined using the partial eta-square. Partial eta-square was represented by the following expression (Norusis/SPSS Inc, 1993).

$$\frac{sums.of.squares.for.the.effect.of.interest}{(sums.of.squares.for.the.effect.of.interst)+(sums.of.squares.for.errors.effect)}$$

For the purpose of this study, it was assumed that interactions among variables would be of practical importance if the partial eta-square was roughly 0.15 or greater. Adequacy of this cutoff value for partial-eta square was supported by the large variation of simple main effects of each variable. Once practically important higher order interactions were identified, simple main effects of those factors at specific levels of other variables were calculated. For a detail discussion of these simple main effects of those factors see Dawadi, 1998.

Results:

This result section was organized in the following way: first, results from the analysis of Average Ability (θ) were discussed, and then results from the analysis of individual ability (Ability θ_1 and Ability θ_2) were discussed. Analysis of individual ability was further divided by the variable dimensional strength: one of two equally important abilities (DS=50/50), major ability of two unequally important abilities (ability represented by 75 percent of the total number of items), and minor ability of two unequally important abilities (ability represented by 25 percent of the total number of items). Relative Impacts were used for identifying problematic treatment conditions and for developing practice guidelines. Absolute Impacts were used in the ANOVA procedure for identifying the statistically significant contrasts of each of the study variables.

Identifying Problematic Conditions while Estimating Average Ability:

Means (mean_{av}), Standard Deviations (std dev), and Relative Impact (percent_{av}) of average ability (IMPCT_AV) and root mean square error of unidimensional data (RMSE_{1D}) for the 64 treatment conditions are presented in Table 2. Also presented in the table are the number of factors and their correlations from the factor analysis.

Insert Table 2 about here

As presented in Table 2, the mean value, which was the absolute difference between the RMSE of unidimensional and two-dimensional data of IMPCT_AV for the 64 treatment conditions, ranged from a low of 0.001 to a high of 0.367. The standard deviations of IMPCT_AV ranged from 0.001 to 0.06, showing a small variation within specific combinations of conditions.

From analyzing Relative Impacts, it was obvious that the correlation was an important variable for classifying the robust and non-robust treatment conditions of this study. When the correlation (COR) was 0.8, the results were generally robust (Relative Impact < 20 percent) regardless of the levels of other variables used. However, when the correlation was 0.3, the results were robust only when the number categories (CAT) was 3 and the dimensional strength (DS) was 50/50. Forty treatment conditions out of the total of 64 were counted by selecting the following treatment conditions: 1) all treatment conditions with correlations of 0.8 and 2) treatment conditions whose correlations were 0.3, number of response categories were 3, and dimensional strength were 50/50. The relative impacts of 34 out of these 40 treatment conditions (85 percent accuracy rate) were not problematic (impact < 20 percent) to the violation when the estimated ability was compared with the true average ability.

For those treatment conditions with correlations of 0.3, the procedure was not robust when the number of response categories was 5 or 3 and the dimensional strength was 25/75. Twenty-four out of 64 treatment conditions fell under those conditions, and the Relative Impacts of those 20 out of 24 treatment conditions were greater than the threshold of 20 percent (83 percent accurate in detecting problematic treatment conditions). The most problematic conditions were identified when the correlation was 0.3, the number of response categories was

5, and the dimensional strength was 25/75. Within these treatment conditions, the Relative Impact ranged from 36.42 to 116.7 percent, a big impact due to the violation of the assumption.

Proposed Guidelines while Estimating Average Ability. Abiding by the rules described earlier, the following guidelines were developed. When estimating the average ability, practitioners can assume the IRT procedure to be robust to the violation if a single factor solution was obtained from the factor analysis. Thirteen out of 14 treatment conditions (93 out of 100 times) were robust to the violation when only one factor was extracted by the factor analysis.

When two factors were extracted by the factor analysis, the correlation between those factors was evaluated. It was revealed that when the correlation between factors obtained by the factor analysis was greater than 0.4, the treatment conditions were not problematic to the violation. Under this condition, 15 out of 18 treatment conditions (83 percent correct decision) were not problematic to the violation. But, when the correlation was smaller than 0.4, the number of response categories and the dimensional strength were critical to the robustness of the procedure. When the correlation was smaller than 0.4, but the number of response categories was 3 and the dimensional strength was 50/50, six out of eight treatment conditions (75 percent correct decision) were not problematic to the violation. All other remaining treatment conditions (24 out of 64 conditions) were problematic to the violation. These 24 treatment conditions were comprised of DS=25/75, and CAT= 5 and 3. For these treatment conditions, when the correlation was smaller than 0.4, twenty out of 24 conditions (83 percent correct decision) were problematic to the violation.

Effects of Study Variables while Estimating the Average Ability. The highest order significant interaction was a 5-way interaction, CAT x COR x DS x SLP x THR (p=.001), and it was also practically important. The contrasts of each variable at each level of the other variables were calculated aggregating over the variable (TL). The variable TL was involved only in two of the two-way interactions: COR x TL and DS x TL; however, the effect of the TL on Impact was not practically important. The effects of each of the study variables obtained from the ANOVA analysis of Absolute Impacts (IMPCT_AV) were ranked according to the magnitude of their contrasts in the Table 3. For practical purposes it was assumed that any simple main effects contrast of Impacts greater than 0.1 would be of practical importance.

Insert Table 3 about here

From the table it can be seen that, when estimating the average ability, CAT(0.316) had the largest effect on Impact followed by COR(-0.263), DS(-0.182), THR(-0.171) and SLP (0.157). The positive value of contrasts of CAT and SLP indicated that the effect of CAT and SLP on Impact was larger when the number of response categories was 5 and when the item slope was 1.0 compared to when the number of response categories was 3, and when the item slope was 0.5, respectively. The negative value of contrasts of COR, DS, and THR indicated that the effect of COR, DS, and THR on Impact was larger when the correlation between two abilities was 0.3, when the dimensional strength was 25/75, and when the range of item category thresholds was -1 to + 1 compared to when the correlation was 0.5, when the dimensional strength was 50/50, and when the range of item category thresholds was -2 to +2, respectively.

For the average ability, the effects of COR, DS, THR and SLP on Impacts were large when the number of response categories was 5. Similarly, the effect of CAT, DS, THR and SLP on Impacts was large when the COR was 0.3. There was no obvious problematic level of the dimensional strength, item slopes, and range of thresholds.

Estimating Individual Ability (θ_1 and θ_2):

While estimating two individual abilities, results were analyzed for three different conditions: 1) when both θ_1 and θ_2 were equally important abilities i.e., when both abilities were represented by an equal number of items (DS=50/50); 2) when ability θ_2 was a major ability represented by 75 percent of the total number of items (DS=25/75); and 3) when the ability θ_1 was a minor ability represented by 25 percent of the total number of items (DS=25/75). Readers are again reminded that PARSCALE did not estimate either the major ability (θ_2) of two unequally important abilities or the minor ability (θ_1) of two unequally important abilities. However, for the ease of discussion in this section, the ability estimate obtained from PARSCALE was assumed to measure one of two equally important abilities, the major ability (θ_2) of two unequally important abilities, and the minor ability (θ_1) of two unequally important abilities.

Means, standard deviations, and percent of IMPCT_1 for ability θ_1 and IMPCT_2 for ability θ_2 are presented in Table 4. Means are Absolute Impacts and percentages are Relative Impacts. Also presented in the table are RMSE of unidimensional data and the results from factor analysis.

Insert Table 4 about here

Estimating One of Two Equally Important Abilities:

There were 32 treatment conditions with equal dimensional strength (50/50). For these conditions Absolute Impact varied from 0.072 to 0.510 for ability one and from 0.082 to 0.76 for ability two. The standard deviations of these conditions ranged from 0.001 to 0.472 for ability one and from 0.001 to 0.452 for ability two. Similarly, the Relative Impact varied from 12 to 224 percent for ability one and from 12 to 270 percent for ability two.

Identifying problematic conditions using study variables. While estimating either one of the two equally important abilities, θ_1 or θ_2, it was found that the IRT procedure was not robust when the correlation was 0.3. The procedure was found most problematic when the correlation of 0.3 was combined with the number of categories of 5. Under these conditions, the Impact for both abilities ranged approximately from 45 to 225 percent. Impacts were lowered when the correlation was increased from 0.3 to 0.8. The combination of the number of categories and the correlation had considerable effect on the Impact. Eight out of eight treatment conditions were problematic to the violation when the correlation of 0.8 was combined with the number of response categories of 5; however, when the same correlation of 0.8 was combined with the number of response categories of 3, only four out of eight treatment conditions were problematic to the violation. It was an improvement of 50 percent when compared with the number of categories of 5. In general, when looking at the bigger picture of estimating either one of the two equally important abilities, it was safe to conclude that procedure for all 32 treatment conditions was not robust to the violation. Fifty-eight out of 64 Impacts were beyond the threshold of 20 percent (91 percent correct decision).

Identifying problematic conditions using proposed practice guidelines. In this study there were 32 treatment conditions with two equally important abilities, resulting in a total of 64 abilities (32 x 2 = 64). When estimating one of two equally important abilities, and when there was only one factor obtained from the factor analysis result, five out of six treatment conditions were problematic to the violation.

When two factors were extracted from the factor analysis, the correlation between those factors was investigated. It was found that no matter what correlation was obtained ($r< 0.74$ and $r > 0.001$), treatment conditions were problematic to the violation. For the two-factor result, 24 out of 26 treatment conditions were problematic to the violation (92 out of 100 times accurate).

Estimating The Major Ability of Two Unequally Important Abilities:

The ability θ_2 was a major ability represented by 75 percent of the total number of items when the dimensional strength was 25/75. The Impact for the estimated major ability of two unequally important abilities was represented by IMPCT_2 in Table 4. The Absolute Impact (mean_2) ranged from 0.008 to 0.106; its standard deviation ranged from 0.0001 to 0.042; and the Relative Impact (percent_2) ranged from 1.35 to 26.15. The range of both Absolute and Relative Impacts showed that the Impacts for the estimated major ability were minimal. As a reminder, any treatment conditions with Relative Impact smaller than 20 percent was considered not problematic to the violation.

Identifying problematic conditions using study variables. While estimating major ability, 28 out of 32 treatment conditions were not problematic (Impact < 20 percent) to the violation. Thus, when estimating the major ability in two unequally important abilities, the procedure under those given treatment conditions was generally robust. These results were correctly identified 88 out of 100 times. There was an exception when the number of response categories was 5 and the correlation was 0.3. Even with this exception, the procedure was robust for 5 out of 8 times.

Thus, in general, we can conclude that when the dimensional strength was 25/75 and the ability being estimated was represented by a larger number of items, the IRT model would be robust to the violation regardless of the level of other variables used.

Identifying problematic conditions using proposed practice guidelines. When one factor was extracted by the factor analysis while estimating the major ability of two unequally important abilities, the treatment conditions were not problematic to the violation. Eight out of eight treatment conditions were identified as not problematic under these conditions. When two factors were extracted by the factor analysis, most of the time the procedure was still robust, i.e., treatment conditions were not problematic to the violation. Under these situations, 19 out of 24 treatment conditions (79 percent correct decision) were not problematic to the violation. All five treatment conditions that were problematic had a correlation of 0.3, and three out of those five conditions had CAT=5.

Estimating The Minor Ability of Two Unequally Important Abilities:

The ability θ_1 was a minor ability represented by 25 percent of the total number of items when the dimensional strength was 25/75. There were 32 treatment conditions when the level of the dimensional strength was fixed to 25/75. The Impact for the estimated minor ability of two unequally important abilities was represented by IMPCT_1 in Table 4. The Absolute Impact (mean_1) ranged from 0.125 to 0.932, and its standard deviation ranged from 0 to 0.067. The Relative Impact (percent_1) ranged from 19 to 439 percent. The range of both Impacts showed a large variation for the estimated minor ability.

Identifying problematic conditions using study variables. Generally, the procedure was not robust (Impact > 20 percent) to the violation while estimating the minor ability of two unequally important abilities. Thirty-one out of 32 treatment conditions were not robust (97 out of 100 times the result was correct). The most problematic treatment conditions occurred when the correlation was 0.3 and the number of response categories was 5. Although the procedure was not robust for all 32 treatment conditions (Impact > 20 percent), the magnitude of violation was much smaller when the correlation was 0.8 and the number of response categories was 3.

Identifying problematic conditions using proposed practice guidelines. When one factor was extracted by the factor analysis, the procedure for seven out of eight treatment conditions (88 percent correct) was not robust. When two factors were extracted, the IRT procedure for all 32 treatment conditions was not robust. Thus, when minor ability was estimated, the procedure was not robust to the violation of the assumption no matter whether one or two factors were extracted by the factor analysis.

Effects of Study Factors for Ability θ_1 and Ability θ_2:

For IMPCT_1 there were three two-way interactions that were significant and practically important: CATxDS, CORxDS, DSxTHR. No interactions with the variables SLP and TL were practically important for IMPCT_1. For IMPACT_2 three two-way interactions that were significant and practically important were CATxDS, CORxDS and DSxSLP. No interactions with variables THR and TL were practically important for IMPACT_2. The effects of each of the study variables according to the size of their contrast of average Impacts were ranked in Table 5. For practical purposes it was assumed that any simple main effects contrast of Impacts greater than 0.1 would be of practical importance.

Insert Table 5 about here

From the table it can be seen that COR (-0.465) had the largest effect on Impact followed by DS (-0.344), THR (-0.149), and CAT(0.145), when estimating the minor ability θ_1. The variable DS (0.308) had the largest effect on Impact followed by COR (-0.248) when estimating the major ability θ_2. The two strongest effects were obtained from the variables COR and DS, with somewhat weaker effects from the variables THR and CAT when estimating the minor ability. Effects were obtained only from the factor DS and COR when estimating the major ability. Two factors SLP and TL did not have any practically important effects on Impact when estimating both individual abilities, θ_1 and θ_2.

The effect of DS on Impact was large for ability θ_1 and ability θ_2 when the level of DS was 25/75 and 50/50, respectively. Thus, depending on the ability estimated, the effect of the variable dimensional strength on Impact was different. While estimating ability θ_1, the effect of COR, DS and THR on the average Impact was always larger when the levels of these variables were 0.3, 25/75, -1 to +1, respectively. However, the effect of CAT on the average Impact was large only when the number of response category was 5.

The negative contrasts of COR, DS, and THR indicated that changing the level of each of these variables from a lower to a higher level, i.e., 0.3 to 0.8, 25/75 to 50/50 and -1 to +1 to -2 to +2, respectively, decreased the effects of Impacts on the estimated ability θ_1. However, for the contrast of CAT, changing the number of response categories from 5 to 3 (higher to lower level), decreased the effects of Impacts on the estimated ability θ_1.

Summary and Conclusion

The primary purpose of this study was to study the robustness of the IRT generalized partial credit (GPC) model to violation of the unidimensionality assumption, and the secondary purpose was to provide guidelines to practitioners using the results from the factor analysis.

Estimating Average Ability:

From the analysis conducted in the first part, i.e., estimating the average of the two true abilities, it was found that regardless of the level of other variables used, the correlation was an important factor in the robustness of the IRT procedure when estimating the average ability. When the true correlation was 0.8, regardless of the other variables used, the results were generally robust to the violation of the assumption. When the true correlation was 0.3, the results were robust only when the number of response categories was 3 and the dimensional strength was 50/50. Treatment conditions were most problematic to the violation when the correlation was 0.3, category was 5, and dimensional strength was 25/75. Thus, from these results it can be seen that the correlation was an important variable followed by the number of response categories and then the dimensional strength.

Drasgow and Parsons (1985) concluded that the unidimensional IRT model was robust to violation of the unidimensionality assumption when the correlation between common factors was 0.4 or higher. This was the same correlation value obtained from the factor analysis result (not true correlation) that was recommended in this study. As a caveat, the Drasgow and Parsons's study was conducted with the dichotomous IRT model and the conditions were different than those used in this study.

` Results from the effects of study variables. Results from the ANOVA analysis were obtained by analyzing Absolute Impacts, and these results were consistent with the results obtained from analyzing the Relative Impacts. The contrasts of the variable category, correlation, and dimensional strength were statistically significant and practically important while estimating the average ability.

The Impacts, or the errors, of the variable category were always large when the number of response categories was increased from 3 to 5, when the correlation was changed from 0.3 to 0.8, and when the dimensional strength was changed from 25/75 to 50/50. Some of the Impacts due to the variables of item slopes, range of thresholds, and test length were statistically significant, but not always practically important. The contrasts ranked by their magnitude when estimating the average ability were as follows: CAT (0.316), which was the biggest contrast, followed by COR (-0.263), DS (-0.181), THR (-0.171), and SLP (-0.104). From these ANOVA results it was concluded that those variables influenced the IRT procedure of estimation of ability when the assumption was violated.

Results From Estimating Single Ability:

Results from the analysis of the Relative Impact were summarized by the variable dimensional strength (DS). In this way it was possible to separate results according to one of two equally important abilities (DS=50/50), major ability of two unequally important abilities (ability represented by 75 percent of the total number of items), and minor ability of two unequally important abilities (ability represented by 25 percent of the total number of items).

When estimating one of the two equally important abilities (DS= 50/50), the procedure was not robust for correlation of 0.3. The procedure was also not robust when the correlation of 0.3 was combined with the number of response categories of 5. When the correlation was increased to 0.8, the procedure was robust only when the number of response categories was 3; however, the procedure was not robust when the number of response categories was 5 and when the correlation was 0.8.

DeAyala (1995) investigated the influence of dimensionality on parameter estimation when each dimension was represented by an equal number of items. When the assumption was violated, DeAyala found that estimated ability parameters were closer to the averages of the true abilities than either of the individual abilities. This result was consistent with the result of this study. In this study when each dimension was represented by an equal number of items, estimated ability was not closer to either of two equally important abilities. Fifty-eight out of 64 Impacts were beyond the threshold of 20 percent (91 percent correct decision).

When estimating a major ability of two unequally important abilities (the ability was represented by 75 percent of the total number of items) results were robust to the violation, regardless of the correlation and the number of response categories used in this study. Results of this study were comparable with the results obtained by Way, Ansley, and Forsyth (1988). These authors found that, as the correlation between two dimensions decreased (from 0.9 to 0.6 to 0.3), estimated ability was strongly correlated with the major (dominant) ability. In the generated data the dominant ability was defined by its higher discrimination value. Also, Folk and Green (1989) found that, as the correlation between two abilities decreased (the correlations used were 1.0, 0.8, 0.6, 0.4 and 0.2), estimated ability was closer to either one of the two abilities.

When estimating a minor ability of two unequally important abilities (ability was represented by 25 percent of the total number of items), generally the IRT procedure was <u>not</u> robust to the violation. The same arguments used earlier from the study of Way, Ansley, and Forsyth (1988) could be used here. Way et. al. (1988) found that, when the correlation between two abilities decreased, the estimated ability was strongly correlated with the dominant ability. A similar conclusion was obtained by Folk and Green (1989) in their study.

Results from the effects of study variables. The results from the ANOVA analysis were helpful in identifying if the contrast of average impacts were meaningful or not. Results from the ANOVA analysis obtained by analyzing Absolute Impacts were consistent with the results obtained from analyzing the Relative Impacts. While estimating the major ability, contrasts of the variable correlation and dimensional strength were practically important, whereas, while estimating the minor ability, the contrasts of the variables category, correlation, dimensional strength, and threshold were practically important. The contrasts of the item slope and test length were not practically important when estimating both major and minor ability.

Guidelines for Practitioners:

The results obtained in this study were derived from the simulated data where true parameters were known. However, in practice these true parameters are not known. Thus, the same simulated data were analyzed using the Principal Component Analysis (PCA) and its results were presented to provide practitioners with some practical guidelines. The guidelines are based on the number of factors extracted, correlation between those factors, dimensional strength, and number of response categories.

During the estimation of the average ability, when a single factor solution was obtained from the PCA, practitioners could assume the IRT procedure to be robust to the violation. When two factors were extracted by the PCA, the correlation between those factors was to be evaluated. When the correlation between factors extracted by the PCA was greater than 0.4, the treatment conditions were not problematic to the violation. Only when the correlation was smaller than 0.4, the variables "number of response categories" and "dimensional strength" were critical to the robustness of the procedure. When the correlation was smaller than 0.4, but the number of response categories was 3 and the dimensional strength was 50/50, practitioners could consider those treatment conditions not problematic to the violation. All other treatment conditions produced by the combination of the number of response categories of 3 and 5, the dimensional strength of 25/75, and correlations of smaller than 0.4 were problematic to the violation.

Those practitioners who would like to use a polytomous IRT model to analyze their two-dimensional data could do so if the correlation between two abilities obtained through the PCA was 0.4 or higher. The caveat is that the estimated abilities are the average of those two abilities. For example, if practitioners are interested in testing the knowledge of mathematics, and, if the test is made up of specific domains within the mathematics area such as algebra and geometry, the estimated abilities are the average of the knowledge of algebra and geometry.

During the estimation of one of the two equally important abilities, when there was only one factor obtained from the PCA results, it was recommended to the practitioners that the IRT procedure would not be robust to the violation. When there were two factors extracted by the PCA, no matter what the correlation was between those factors $r > 0.001$ and $r < 0.74$), all treatment conditions were problematic to the violation.

During the estimation of the major ability of two unequally important abilities, when there was only one factor extracted by the PCA, it was recommended that under those conditions the treatment conditions were <u>not</u> problematic to the violation, i.e., the IRT procedure was robust to the violation. When there were two factors extracted by the PCA, generally the procedure was still robust. Thus, when a major ability was being estimated, practitioners could use the IRT procedure because it was robust to the violation of the assumption under those given treatment conditions.

During the estimation of the minor ability of two unequally important abilities, when one factor was extracted by the PCA, it was recommended to practitioners that the procedure was not robust under any treatment conditions. When two factors were extracted, the IRT procedure for all treatment conditions was also not robust. Thus, when a minor ability was being estimated, no matter whether one or two factors were extracted by the PCA, practitioners should not use the IRT procedure because it was not robust to the violation of the assumption under those given treatment conditions.

This was a computer simulation study and results obtained should not be generalized beyond the scope of parameters simulated in this study. All recommendations provided in this study were also limited to the scope of parameters of this study. This study included only two distinct dimensions when generating data to violate unidimensionality. Also, all variables (parameters) that were varied in this study were fixed to two levels. Variables (parameters) such as dimensional strength, correlation, and number of response categories were sensitive to the violation. Thus, conducting studies by varying the level of those sensitive variables may provide some definite answers to the question, under which conditions the IRT model would be robust. The results of this study showed that the procedure was robust to the violation for the correlation of 0.3 when the number of response categories was 3 and the dimensional strength was 50/50. However, further investigation could be conducted to determine if either the number of response categories or the dimensional strength was critical for results to be robust when the correlation was small.

Table 1

Variables and Their Levels.

Description of Variables	Variable Label	Levels of Variable	
Category	CAT	3	5
Correlation	COR	0.3	0.8
Dimensional Strength	DS	25/75	50/50
Slope	SLP	0.5	1.0
Threshold	THR	-1 to +1	-2 to +2
Test Length	TL	8	16

CAT: number of response categories for each item in the test

COR: correlation between two abilities in two-dimensional data; for unidimensional data COR=1.0

DS: dimensional strength (or relative importance) determined by the number of items in each dimension in the test. For example, 25/75 represents 25 percent of the total number of items in a test is represented by one dimension, and the remaining 75 percent of the total number of items in the test is represented by the other dimension.

SLP: slope of items used in the test (a concept similar to discrimination in dichotomous IRT models).

THR: the range of values of item thresholds (the lowest and the highest). The number of thresholds depends on the number of item categories.

TL: number of items used (8 and 16 items) in the test.

Table 2
Mean, Standard Deviation and Percent of IMCT_AV and Factor Analysis Results

VARIABLES AND THEIR LEVELS						$RMSE_{ID}$	IMPCT_AV			Factor Analysis Results	
cat	cor	ds	slp	thr	tl		$mean_{av}$	std dev	$percent_{av}$	factor	corr
3	0.3	25/75	0.5	-1to+1	8	0.591	0.099	0.061	16.75	2	.12515
3	0.3	25/75	0.5	-1to+1	16	0.439	0.193	0.015	43.96	2	.17537
3	0.3	25/75	0.5	-2to+2	8	0.6636	0.072	0.021	10.85	2	.12082
3	0.3	25/75	0.5	-2to+2	16	0.5146	0.112	0.017	21.77	2	.16869
3	0.3	25/75	1.0	-1to+1	8	0.395	0.226	0.01	57.22	2	.18028
3	0.3	25/75	1.0	-1to+1	16	0.2945	0.273	0.003	92.7	2	.22355
3	0.3	25/75	1.0	-2to+2	8	0.5946	0.073	0.001	12.28	2	.09725
3	0.3	25/75	1.0	-2to+2	16	0.4363	0.123	0.012	28.19	2	.20639
3	0.3	50/50	0.5	-1to+1	8	0.5726	0.060	0.015	10.48	2	.15981
3	0.3	50/50	0.5	-1to+1	16	0.4403	0.079	0.006	17.94	2	.17791
3	0.3	50/50	0.5	-2to+2	8	0.6667	0.033	0.027	4.95	2	.11369
3	0.3	50/50	0.5	-2to+2	16	0.5	0.062	0.018	12.4	2	.19552
3	0.3	50/50	1.0	-1to+1	8	0.3925	0.084	0.012	21.4	2	.21080
3	0.3	50/50	1.0	-1to+1	16	0.2941	0.075	0.002	25.5	2	.25287
3	0.3	50/50	1.0	-2to+2	8	0.5737	0.030	0.031	5.23	2	.15253
3	0.3	50/50	1.0	-2to+2	16	0.4325	0.066	0.002	15.26	2	.20765
3	0.8	25/75	0.5	-1to+1	8	0.591	0.005	0.054	0.85	1	.00000
3	0.8	25/75	0.5	-1to+1	16	0.439	0.020	0.037	4.56	2	.46220
3	0.8	25/75	0.5	-2to+2	8	0.6636	0.001	0.027	0.15	1	.00000

Table 2-- Continued

VARIABLES AND THEIR LEVELS						RMSE$_{ID}$	IMPCT_AV			Factor Analysis Results	
cat	cor	ds	slp	thr	tl		mean$_{av}$	std dev	percent$_{av}$	factor	corr
3	0.8	25/75	0.5	-2to+2	16	0.5146	0.007	0.009	1.36	2	.42022
3	0.8	25/75	1.0	-1to+1	8	0.395	0.049	0.027	12.41	1	.00000
3	0.8	25/75	1.0	-1to+1	16	0.2945	0.048	0.004	16.3	2	.63519
3	0.8	25/75	1.0	-2to+2	8	0.5946	-0.015	0.02	-2.52	1	.00000
3	0.8	25/75	1.0	-2to+2	16	0.4363	0.019	0.003	4.35	2	.48408
3	0.8	50/50	0.5	-1to+1	8	0.5726	0.001	0.008	0.17	2	.39910
3	0.8	50/50	0.5	-1to+1	16	0.4403	-0.011	0.002	-2.5	2	-.50917
3	0.8	50/50	0.5	-2to+2	8	0.6667	0.010	0.059	1.5	1	.00000
3	0.8	50/50	0.5	-2to+2	16	0.5	-0.002	0.026	-0.4	2	-.44047
3	0.8	50/50	1.0	-1to+1	8	0.3925	0.007	0.017	1.78	1	.00000
3	0.8	50/50	1.0	-1to+1	16	0.2941	0.007	0.006	2.38	2	.68043
3	0.8	50/50	1.0	-2to+2	8	0.5737	-0.004	0.003	-0.7	2	.40260
3	0.8	50/50	1.0	-2to+2	16	0.4325	0.006	0.01	1.39	2	-.54988
5	0.3	25/75	0.5	-1to+1	8	0.3977	0.257	0.004	64.62	2	.20519
5	0.3	25/75	0.5	-1to+1	16	0.2988	0.297	0.011	99.4	2	.21912
5	0.3	25/75	0.5	-2to+2	8	0.4393	0.160	0.001	36.42	2	.19816
5	0.3	25/75	0.5	-2to+2	16	0.336	0.218	0.004	64.88	2	.20738
5	0.3	25/75	1.0	-1to+1	8	0.2836	0.331	0.002	116.7	2	.24249
5	0.3	25/75	1.0	-1to+1	16	0.2122	0.367	0.007	173	2	.27002

Table 2--continued

VARIABLES AND THEIR LEVELS						$RMSE_{1D}$	IMPCT_AV			Factor Analysis Results	
cat	cor	ds	slp	thr	tl		$mean_{av}$	std dev	$percent_{av}$	factor	corr
5	0.3	25/75	1.0	-2to+2	8	0.3307	0.244	0.005	73.78	2	.20871
5	0.3	25/75	1.0	-2to+2	16	0.2447	0.281	0.01	114.8	2	.25707
5	0.3	50/50	0.5	-1to+1	8	0.4142	0.097	0.033	23.42	2	.23212
5	0.3	50/50	0.5	-1to+1	16	0.3022	0.094	0.013	31.11	2	.23570
5	0.3	50/50	0.5	-2to+2	8	0.4564	0.058	0.015	12.71	2	.18190
5	0.3	50/50	0.5	-2to+2	16	0.34	0.071	0.021	20.88	2	.23673
5	0.3	50/50	1.0	-1to+1	8	0.2812	0.246	0.013	87.48	2	.00113
5	0.3	50/50	1.0	-1to+1	16	0.2075	0.257	0.008	123.9	2	.00062
5	0.3	50/50	1.0	-2to+2	8	0.3308	0.078	0.006	23.58	2	.00162
5	0.3	50/50	1.0	-2to+2	16	0.2394	0.084	0.004	35.09	2	.26144
5	0.8	25/75	0.5	-1to+1	8	0.3976	0.050	0.041	12.58	1	.00000
5	0.8	25/75	0.5	-1to+1	16	0.2988	0.061	0.014	20.41	2	.60888
5	0.8	25/75	0.5	-2to+2	8	0.4393	0.040	0.033	9.11	1	.00000
5	0.8	25/75	0.5	-2to+2	16	0.336	0.034	0.007	10.12	2	.56406
5	0.8	25/75	1.0	-1to+1	8	0.2836	0.077	0.031	27.15	1	.00000
5	0.8	25/75	1.0	-1to+1	16	0.2122	0.095	0.006	44.77	2	.72349
5	0.8	25/75	1.0	-2to+2	8	0.3307	0.047	0.001	14.21	1	.00000
5	0.8	25/75	1.0	-2to+2	16	0.2447	0.061	0.028	24.93	2	.68542
5	0.8	50/50	0.5	-1to+1	8	0.4142	-0.011	0.008	-2.66	1	.00000

Table 2--continued

VARIABLES AND THEIR LEVELS						$RMSE_{ID}$	IMPCT_AV			Factor Analysis Results	
cat	cor	ds	slp	thr	tl		$mean_{av}$	std dev	$percent_{av}$	factor	corr
5	0.8	50/50	0.5	-1to+1	16	0.3022	0.009	0.007	2.98	2	.67078
5	0.8	50/50	0.5	-2to+2	8	0.4564	0.020	0.024	4.38	1	.00000
5	0.8	50/50	0.5	-2to+2	16	0.34	0.008	0.022	2.35	2	-.61329
5	0.8	50/50	1.0	-1to+1	8	0.2812	0.006	0.012	2.13	1	.00000
5	0.8	50/50	1.0	-1to+1	16	0.2075	0.011	0.006	5.3	2	.74360
5	0.8	50/50	1.0	-2to+2	8	0.3308	0.004	0.001	1.21	1	.00000
5	0.8	50/50	1.0	-2to+2	16	0.2394	0.006	0.008	2.51	2	.69518

cat no. of response categories of each item in the test.
cor correlation between two abilities in two-dimensional data; correlation for unidimensional data=1.00
ds dimensional strength determined by the number of items in each dimension in the test. For example, in a test with the ds of 25/75, 25 percent of the total number of items represent dimension one and remaining 75 percent of the total number of items represent dimension two.
slp slope of each item in the test.
tl test length (8 items and 16 items tests).
percent mean / $(RMSE)_{ID}$ x 100
std dev std. dev. of those two means
factor no of factors extracted from the original two-dimensional data.
corr correlation of those extracted factors by the factor analysis.

thr the range of item thresholds (the lowest and the highest).
$(RMSE)_{ID}$ $[(\hat{\theta} - \theta)/n]^{1/2}$ for unidimensional data
$mean_{av}$ average of two replications in each cell
Factor Analysis Results results obtained from conducting PCA.

Table 3
Results of the ANOVA analysis of Absolute Impact while estimating Average Ability

Effects of	Biggest Contrast	Conditions (Variables and their Levels)				
		CAT	COR	DS	SLP	THR
CAT	0.316	-	0.3	25/75	0.5	-1 to +1
COR	-0.263	5	-	25/75	1.0	-1 to +1
COR	-0.244	5	-	50/50	1.0	-1 to +1
COR	-0.222	5	-	25/75	0.5	-1 to +1
COR	-0.208	5	-	25/75	1.0	-2 to +2
COR	-0.202	3	-	25/75	1.0	-1 to +1
DS	-0.182	5	0.3	-	0.5	-1 to +1
DS	-0.181	5	0.3	-	1.0	-2 to +2
THR	-0.171	5	0.3	50/50	1.0	-
SLP	0.157	5	0.3	50/50	-	-1 to +1

Table 4
Mean, Standard Deviation and Percent of IMPCT_1 and IMPCT_2 and Factor Analysis Results

VARIABLES AND THEIR LEVELS						RMSE$_{1D}$	IMPCT_1			IMPCT_2			Factor Analysis[#]	
cat	cor	ds	slp	thr	tl		mean$_1$	std dev	percent$_1$	mean$_2$	std dev	percent$_2$	factor	Corr
3	0.3	25/75	0.5	-1 to 1	8	0.591	0.530	0.063	89.68	0.037	0.042	6.26	2	.12515
3	0.3	25/75	0.5	-1 to 1	16	0.439	0.682	0.021	155.35	0.052	0.002	11.85	2	.17537
3	0.3	25/75	0.5	-2 to 2	8	0.6636	0.448	0.008	67.51	0.076	0.038	11.45	2	.12082
3	0.3	25/75	0.5	-2 to 2	16	0.5146	0.534	0.033	103.77	0.106	0.025	20.60	2	.16869
3	0.3	25/75	1.0	-1 to 1	8	0.395	0.734	0.011	185.82	0.051	0.014	12.91	2	.18028
3	0.3	25/75	1.0	-1 to 1	16	0.2945	0.814	0.002	276.4	0.047	0.007	15.96	2	.22355
3	0.3	25/75	1.0	-2 to 2	8	0.5946	0.486	0.011	81.74	0.058	0.022	9.75	2	.09725
3	0.3	25/75	1.0	-2 to 2	16	0.4363	0.584	0.023	133.85	0.096	0.01	22.00	2	.20639
3	0.8	25/75	0.5	-1 to 1	8	0.591	0.151	0.067	25.55	0.008	0.033	1.35	1	.00000
3	0.8	25/75	0.5	-1 to 1	16	0.439	0.200	0.033	45.56	0.023	0.035	5.24	2	.16220
3	0.8	25/75	0.5	-2 to 2	8	0.6636	0.125	0.028	18.84	0.015	0.024	2.26	1	.00000
3	0.8	25/75	0.5	-2 to 2	16	0.5146	0.170	0.007	33.04	0.009	0.01	1.75	2	.42022
3	0.8	25/75	1.0	-1 to 1	8	0.395	0.240	0.029	60.76	0.041	0.021	10.38	1	.00000
3	0.8	25/75	1.0	-1 to 1	16	0.2945	0.282	0.004	95.76	0.024	0.005	8.15	2	.63519
3	0.8	25/75	1.0	-2 to 2	8	0.5946	0.132	0.029	22.20	-0.008	0.005	-1.34	1	.00000
3	0.8	25/75	1.0	-2 to 2	16	0.4363	0.202	0.006	46.30	0.018	0.005	4.13	2	.48408
5	0.3	25/75	0.5	-1 to 1	8	0.3977	0.765	0.004	192.36	0.057	0.015	14.33	2	.20519
5	0.3	25/75	0.5	-1 to 1	16	0.2988	0.842	0.012	281.79	0.033	0.000	11.04	2	.21912

Table 4--continued

VARIABLES AND THEIR LEVELS						RMSE$_{1D}$	IMPCT_1			IMPCT_2			Factor Analysis[#]	
cat	cor	ds	slp	thr	tl		mean$_1$	std dev	percent$_1$	mean$_2$	std dev	percent$_2$	factor	Corr
5	0.3	25/75	0.5	-2 to 2	8	0.4393	0.626	0.011	142.50	0.093	0.002	21.17	2	.19816
5	0.3	25/75	0.5	-2 to 2	16	0.336	0.735	0.000	218.75	0.072	0.019	21.43	2	.20738
5	0.3	25/75	1.0	-1 to 1	8	0.2836	0.880	0.001	310.30	0.034	0.000	11.99	2	.24249
5	0.3	25/75	1.0	-1 to 1	16	0.2122	0.932	0.007	439.21	0.035	0.000	16.49	2	.27002
5	0.3	25/75	1.0	-2 to 2	8	0.3307	0.771	0.008	233.14	0.051	0.006	15.42	2	.20871
5	0.3	25/75	1.0	-2 to 2	16	0.2447	0.831	0.007	339.60	0.064	0.029	26.15	2	.25707
5	0.8	25/75	0.5	-1 to 1	8	0.3976	0.243	0.043	61.12	0.038	0.029	9.56	1	.00000
5	0.8	25/75	0.5	-1 to 1	16	0.2988	0.291	0.017	97.39	0.034	0.000	11.38	2	.60888
5	0.8	25/75	0.5	-2 to 2	8	0.4393	0.216	0.03	49.17	0.041	0.028	9.33	1	.00000
5	0.8	25/75	0.5	-2 to 2	16	0.336	0.245	0.000	72.92	0.034	0.017	10.12	2	.56406
5	0.8	25/75	1.0	-1 to 1	8	0.2836	0.318	0.03	112.13	0.029	0.018	10.23	1	.00000
5	0.8	25/75	1.0	-1 to 1	16	0.2122	0.362	0.01	170.59	0.029	0.007	13.67	2	.72349
5	0.8	25/75	1.0	-2 to 2	8	0.3307	0.261	0.000	78.92	0.037	0.002	11.19	1	.00000
5	0.8	25/75	1.0	-2 to 2	16	0.2447	0.310	0.03	126.69	0.037	0.018	15.12	2	.68542
3	0.3	50/50	0.5	-1 to 1	8	0.5726	0.272	0.011	47.50	0.315	0.019	55.01	2	.15981
3	0.3	50/50	0.5	-1 to 1	16	0.4403	0.378	0.005	85.85	0.315	0.008	71.54	2	.17791
3	0.3	50/50	0.5	-2 to 2	8	0.6667	0.254	0.042	38.10	0.245	0.017	36.75	2	.11369
3	0.3	50/50	0.5	-2 to 2	16	0.5	0.313	0.039	62.60	0.32	0.000	64.00	2	.19552
3	0.3	50/50	1.0	-1 to 1	8	0.3925	0.385	0.018	98.09	0.35	0.001	89.17	2	.2108

Table 4--continued

VARIABLES AND THEIR LEVELS						$RMSE_{ID}$	IMPCT_1			IMPCT_2			Factor Analysis[#]	
cat	cor	ds	slp	thr	tl		$mean_1$	std dev	$percent_1$	$mean_2$	std dev	$percent_2$	factor	Corr
3	0.3	50/50	1.0	-1 to 1	16	0.2941	0.417	0.004	141.79	0.389	0.001	132.27	2	.25287
3	0.3	50/50	1.0	-2 to 2	8	0.5737	0.280	0.005	48.81	0.263	0.046	45.84	2	.15253
3	0.3	50/50	1.0	-2 to 2	16	0.4325	0.331	0.007	76.53	0.350	0.006	80.92	2	.20765
3	0.8	50/50	0.5	-1 to 1	8	0.5726	0.089	0.016	15.54	0.075	0.001	13.10	2	.39910
3	0.8	50/50	0.5	-1 to 1	16	0.4403	0.092	0.001	20.89	0.094	0.007	21.35	2	-.50917
3	0.8	50/50	0.5	-2 to 2	8	0.6667	0.079	0.048	11.85	0.082	0.066	12.30	1	.00000
3	0.8	50/50	0.5	-2 to 2	16	0.5	0.079	0.038	15.80	0.100	0.014	20.00	2	-.44047
3	0.8	50/50	1.0	-1 to 1	8	0.3925	0.105	0.019	26.75	0.129	0.011	32.87	1	.00000
3	0.8	50/50	1.0	-1 to 1	16	0.2941	0.141	0.003	47.94	0.145	0.005	49.30	2	.68043
3	0.8	50/50	1.0	-2 to 2	8	0.5737	0.072	0.014	12.55	0.083	0.01	14.47	2	.40260
3	0.8	50/50	1.0	-2 to 2	16	0.4325	0.106	0.006	24.51	0.110	0.011	25.43	2	-.54988
5	0.3	50/50	0.5	-1 to 1	8	0.4142	0.510	0.066	123.13	0.192	0.037	46.35	2	.23212
5	0.3	50/50	0.5	-1 to 1	16	0.3022	0.376	0.128	124.42	0.431	0.128	142.62	2	.23570
5	0.3	50/50	0.5	-2 to 2	8	0.4564	0.332	0.01	72.74	0.323	0.016	70.77	2	.18190
5	0.3	50/50	0.5	-2 to 2	16	0.34	0.383	0.032	112.65	0.378	0.012	111.18	2	.23673
5	0.3	50/50	1.0	-1 to 1	8	0.2812	0.134	0.011	47.65	0.760	0.013	270.27	2	.00113
5	0.3	50/50	1.0	-1 to 1	16	0.2075	0.466	0.472	224.58	0.474	0.452	228.43	2	.00062
5	0.3	50/50	1.0	-2 to 2	8	0.3308	0.401	0.014	121.22	0.375	0.020	113.36	2	.00162
5	0.3	50/50	1.0	-2 to 2	16	0.2394	0.457	0.021	190.89	0.412	0.013	172.10	2	.26144

Table 4--continued

VARIABLES AND THEIR LEVELS						RMSE_{ID}	IMPCT_1			IMPCT_2			Factor Analysis[#]	
cat	cor	ds	slp	thr	tl		mean_1	std dev	percent_1	mean_2	std dev	percent_2	factor	Corr
5	0.8	50/50	0.5	-1 to 1	8	0.4142	0.092	0.001	22.21	0.104	0.008	25.11	1	.00000
5	0.8	50/50	0.5	-1 to 1	16	0.3022	0.143	0.007	47.32	0.139	0.006	45.99	2	.67078
5	0.8	50/50	0.5	-2 to 2	8	0.4564	0.121	0.004	26.51	0.110	0.039	24.10	1	.00000
5	0.8	50/50	0.5	-2 to 2	16	0.34	0.129	0.028	37.94	0.132	0.016	38.82	2	-.61329
5	0.8	50/50	1.0	-1 to 1	8	0.2812	0.147	0.004	52.28	0.144	0.027	51.21	1	.00000
5	0.8	50/50	1.0	-1 to 1	16	0.2075	0.180	0.006	86.75	0.173	0.006	83.37	2	.74360
5	0.8	50/50	1.0	-2 to 2	8	0.3308	0.129	0.011	38.99	0.130	0.008	39.30	1	.00000
5	0.8	50/50	1.0	-2 to 2	16	0.2394	0.153	0.002	63.91	0.169	0.014	70.59	2	.69518

cat no. of response categories for each item in the test.

cor correlation between two abilities in two-dimensional data; correlation for unidimensional data=1.0.

ds dimensional strength (relative importance) determined by the number of items in each dimension in the test. For example, in a test with the dimensional strength of 25/75, 25 percent of the total number of items represent one dimension and remaining number of items represent the other dimension.

slp slope of each item in the test

tl test length (8 items and 16 items test)

std dev standard deviation of the IMPCT_1 in each cell

corr correlation between two factors extracted by factor analysis

[#] same as presented in Table 4.2

thr the range of item threshold (the lowest and the highest)

$(\text{RMSE})_{\text{ID}} = [(\hat{\theta} - \theta)/n]^{1/2}$ for unidimensional data, n = 1000

percent $= \text{mean}/(\text{RMSE})_{\text{ID}}$

factor no of factors determined from the factor analysis

mean mean of IMPCT_1 in each cell for a fixed set of treatment conditions

Table 5
Results of the ANOVA analysis of Absolute Impact while estimating Ability θ_1 and θ_2.

Effects of	Biggest Contrast When Estimating			
	Ability θ_1	Condition	Ability θ_2	Condition
COR	-0.465	DS=25/75	-	-
DS	-0.344	COR=0.3	0.308	COR=0.3
DS	-0.280	CAT=5	-	-
COR	-0.239	DS=50/50	-0.248	DS=25/75
DS	-0.183	CAT=3	-	-
THR	-0.149	DS=25/75	-	-
CAT	0.145	DS=25/75	-	-

References

Andrich, David. (1978a). A Rating Formulation for Ordered Response Categories. Psychometrika. 43. pp. 561-573.

Andrich, David. (1978b). Application of a Psychometric Rating Model to Ordered Categories Which are Scored with Successive Integers. Applied Psychological Measurement. 2 4 pp.581-594.

Ackerman, T.A. (1989). Unidimensional IRT Calibration of Compensatory and Noncompensatory Multidimensional Items. Applied Psychological Measurement. 13 2, pp. 113-127.

Ansley, T.N. and Forsyth, R.A. (1985). An Examination of the Characteristics of Unidimensional IRT Parameter Estimates Derived From Two-Dimensional Data. Applied Psychological Measurement. 9. 1. pp. 37-48.

California's New Academic Assessment System. (1996, January). National Council on Measurement in Education Quarterly Newsletter, 3, 1.

Crocker, L. (1995, Winter). Editorial. Educational Measurement: Issues and Practice, 14, 4.

Dawadi, B. R. (1998). Robustness of the Polytomous IRT Model to the Violations of the Unidimensionality Assumption. Dissertation. Florida State University, Tallahassee, FL.

DeAyala, R.J. (1995a). The Influence of Dimensionality on Estimation in the Partial Credit Model. Educational and Psychological Measurement. 55 3. pp.407-222.

DeAyala, R.J. (1995b). Item Parameter Recovery for the Nominal Response Model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA April 18-22, 1995.

Dirir, M.A. & Sinclair, N. (April, 1996). On reporting IRT ability scores when the test is not unidimensional. A paper presented at the annual meeting of the NCME, New York.

Dodd, B. G., Koch, W. R., and DeAyala, R. J. (1993). Computerized Adaptive Testing Using the Rasch Partial Credit Model: Effects of Item Pool Characteristics and Different Stopping Rules. Educational and Psychological measurement. 53. pp. 61-77.

Dorans, N. J., & Kingston, N. M. (1985). The Effect of Violations of Unidimensionality on the Estimation of Item and Ability Parameters and on Item Response Theory Equating of the GRE Verbal Scale. Journal of Educational Measurement. 22 (4), 249-262.

Downing, S. M. and Haladyna, T. M. (1996). A Model for Evaluating High-Stakes Testing Programs: Why the Fox Should Not Guard the Chicken Coop. Educational Measurement: Issues and Practice, 15 (1), 5-12.

Drasgow, F. and Parsons, C.K. (1983). Application of Unidimensional Item Response Theory Models to Multidimensional Data. Applied Psychological Measurement. 7. 2. pp. 189-199.

Folk, V.G. and Green, B.F. (1989). Adaptive Estimation When the Unidimensionality Assumption of IRT is Violated. Applied Psychological Measurement, 13 (4), 373-389.

Harris, J, Laan, S. and Mossenson, L. (1988). Applying Partial Credit Analysis to the Construction of Narrative Writing Test. Applied Measurement in Education. 1. 4. pp. 335-346.

Hambleton, R.K. (1989). Principles and Selected Applications of Item Response Theory. In R. Linn (Ed.), Educational Measurement. (3rd ed. pp.147-200). New York:Wiley

Harrison, D.A. (1986). Robustness of IRT Parameter Estimation to Violations of the Unidimensionality Assumption. Journal of Educational Statistics. 11. 2. pp. 91-115.

Hambleton, R. and Swaminathan, H. (1985). Item Response Theory: Principles and Applications. Boston, MA: Kluwer.Nijhoff Publishing.

Kirisci, L. and Hsu, T.C. (1995). The Robustness of BILOG to Violations of Assumptions of Unidimensionality of Test Items and Normality of Ability. Paper presented at the annual meeting of the NCME, San Francisco, April, 1995.

Luecht, R. M. and Miller, T. R. (1992a). Unidimensional Calibration and Interpretation of Composite Traits for Multidimensional Tests. Applied Psychological Measurement. 16. 3, pp. 279-293.

Luecht R. M. and Miller, T. R. (1992b). Consideration of Multidimensionality in Polytomous Item Response Models. Paper presented at the annual meeting of the AERA, San Francisco, CA. April.

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. Psychometrika. 47. 2. pp. 149-174.

Muraki, E. (1993). Information Functions of the Generalized Partial Credit Model. Applied Psychological Measurement. 17,(4). 351-363.

Muraki, E. & Carlson, J. E. (1993). Full-information Factor Analysis for Polytomous Item Responses. Paper presented at the annual meeting of the AERA (Atlanta, GA, April).

Muraki, E. & Carlson, J. E. (1995). Full-information Factor Analysis for Polytomous Item Responses. Applied Psychological Measurement. 19. 1. pp.73-90

Muraki, E.. (1992a). A Generalized Partial Credit Model: Application of an EM Algorithm. Applied Psychological Measurement. 16 2. pp. 159-176.

Muraki, E. & Ankenmann, R. D. (1993). Applying the Generalized Partial Credit Model to Missing Responses: Implementing the Scoring Function and a Lower Asymptote Parameter. A paper presented at the annual meeting of the AERA, Atlanta, GA.

Muraki, E. & Bock, R.D. (1993). PARSCALE: IRT based Test Scoring and Item Analysis for Graded Open-ended Exercises and Performance Tasks. Scientific Software International, Chicago:IL

Muraki, E. (1990). Fitting a Polytomous Item Response Model to Likert-Type Data. Applied Psychological Measurement. 14. 1. pp. 59-71.

Muraki, E. (1996). RESGEN: Item Response Generator. Version 2.0, Educational Testing Service, Princeton, New Jersey.

Norusis, M. J./SPSS Inc. (1993). SPSS Manual. SPSS, Inc. Chicago, IL

Oosterhof, A. C. (1990). Classroom Applications of Educational Measurement. Columbus, OH: Merrill Publishing Company.

Oshima, T.C. and Miller, M.D. (1990). Multidimensionality and IRT-Based Item Invariance Indexes: The Effect of Between-Group Variation in Trait Correlation. Journal of Educational Measurement. 27. 3. pp. 273-283.

Oshima, T.C. and Miller, M.D. (1992). Multidimensionality and Item Bias in Item Response Theory. Applied Psychological Measurement. 16. 3. pp. 237-248.

Reckase, M.D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. Journal of Educational Statistics. 4, 3, pp.207-230.

Reckase, M.D. (1985). The Difficulty of Test Items That Measure More than One Ability.

Applied Psychological Measurement. 9. 4. pp.401-412.

Reckase, M.D., Ackerman, T.A. & Carlson, J.E. (1988). Building Unidimensional Test Using Multidimensional Items. Journal of Educational Measurement. 25. 3. pp.193-203.

Sykes, R.C., Yen, W. & Ito, K. (1996). Scaling Polytomous Items That Have Been Scored by Two Raters. Paper presented at the annual meeting of the NCME (New York, NY April).

Tate, R. L. (1992). Maintaining Scale Consistency for Florida Writing Assessment Programs. Student Assessment Services, Bureau of Education Information and Assessment Services, Department of Education, Tallahassee, Florida.(Unpublished)

Tate, R. L. (1993). Polytomous IRT Scaling of Florida Writing Assessment Data. Student Assessment Services, Bureau of Education Information and Assessment Services, Department of Education, Tallahassee, Florida. (Unpublished)

Traub, R. E. (1983). A Priori Considerations in Choosing an Item Response Model. In R. K. Hambleton (Ed.), Application of item Response Theory (pp. 57-70). Vancouver, British Columbia: Educational Research Institute of British Columbia.

Wainer, H. & Thissen, D. (1987). Estimating Ability with the Wrong Model. Journal of Educational Statistics Winter. 12, 4. pp.339-368.

Way, W.D., Ansley, T.N., & Forsyth, R.A. (1988). The Comparative Effects of Compensatory and Noncompensatory Two-Dimensional Data on Unidimensional IRT Estimates. Applied Psychological Measurement. 12. 3. pp. 239-252.

Wilson, M & Iventosch, L. (1988). Using the Partial Credit Model to Investigate Responses to Structured Sub-Tests. Applied Measurement in Education. 1. 4. pp.319-334.

Wright, B.D., Congdon, R., & Schultz, M. (1989). A user's guide to MSTEPS (version2.4). Chicago: MESA Psychometric Laboratory.

Zeng, Lingjia. (1989). Robustness of Unidimensional Latent Trait Models When Applied to Multidimensional Data. Dissertation. University of Georgia, Athens, GA.

U.S. Depment of Education
Office of Educational Rerch and Improvement (OERI)
National Librarf Education (NLE)
Educational Resources Irmation Center (ERIC)

ERIC

ReproductioRelease

(Specific Documen.

I. DOCUMENT IDENTIFICATION:

Title: Robustness of the Polytomous IRT Model to Violations of the Unidimensionality Assu

Author(s): Bhaskar R. Dawadi, Ph.D.

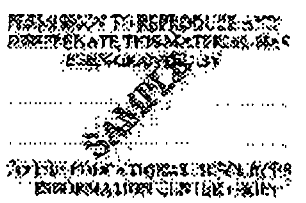
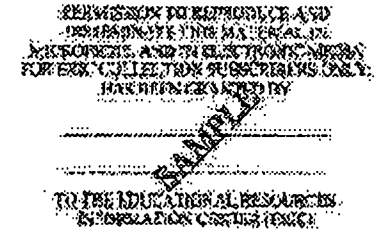
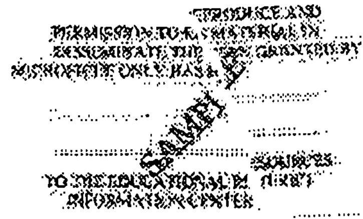
Corporate Source:

blication Date: April 1999

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign in the indicated space following.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all 2B documents Leve
		
Level 1	Level 2A	Level 2B
☑	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g. electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche, or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature: Bhaskar R Dawadi	Printed Name/Position/Title: Bhaskar R. Dawadi, Ph.D., Consultant

Organization/Address:
Georgia Examining Boards
166 Pryor Street, SW, #303
Atlanta, GA 30303

Telephone: (404) 656-3903

Fax: (404) 657-6389

E-mail Address: brdawadi@sos.state.ga.us

Date: 4/10/1999